PREPARING THE DATA

- Two most common data-types:
  * Numeric $\longmapsto$ order relation + distance relation
  * Categorical
    - → if binary: with two values (0 and 1)
    - → Code: 1000, 0100, 0010, 0001

- Missing values, distortions, misrecording, inadequate sampling can be characteristics of data.

Transformations Of Raw Data

→ missing values
→ distortions
→ misrecording
→ inadequate sampling

1- Normalizations

- Based on distance computation between points in an n-dimensional space.
- Scaled into eg., [-1, 1] or [0, 1]
- If not normalized, the distance measures will overweight those features that have, on average, larger values.
  - decimal scaling
  - min-max scaling
  - std normalization

Decimal Scaling:

- Range [-1, 1]
- $v'(i) = v(i)/(10^k)$, for the smallest k such that $\max\{|v'(i)|\} < 1$
- For example, if the largest value in the set is 455 and the smallest value is -834, then the maximum absolute value of the feature becomes .834, and the divisor for all $v(i)$ is 1000 ($k=3$).

-834, ---, 455

(-0.834, ---, 0.455)

Min-max Normalization:

- $V'(i) = [V(i) - \min\{V(i)\}] / [\max\{V(i)\} - \min\{V(i)\}]$

- $[0,1]$ or $[-1,1]$

Std Normalization

- $V'(i) = [V(i) - \mu(V)] / \sigma(V)$
- Transforms data into a form unrecognizable from the original data.
- z-score norm

Example

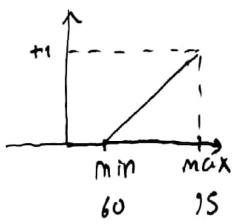
$D = \{70, 90, 85, 95, 70, 90, 60\}$, apply normalization techniques and find the new values of 65 and 85.

a) min-max $\rightarrow [0,1]$

b) min-max $\rightarrow [-1,1]$

c) std normalization

a) max = 95, min = 60

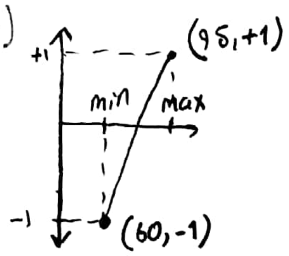


$$V' = \frac{V-60}{95-60} = \frac{V-60}{35}$$

$V = 65 \Rightarrow V' = 5/35$

$V = 85 \Rightarrow V' = 25/35$

b)



$$\frac{y_2 - y_1}{x_2 - x_1} = \frac{y - y_1}{x - x_1} \quad \text{solve}$$

$$\frac{2}{35} = \frac{y-1}{x-95}$$

$$\frac{2x - 190}{35} + 1 = y$$

c) $m = 80$

$S = 13.22$

$$V' = \frac{V-80}{13.22}$$

$V = 65 \Rightarrow V' = -1.23$

$V = 85 \Rightarrow V' = 0.41$

$$S = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

2- Data Smoothing

- May be considered as random variations of the same underlying value.

- Bining
  - Sort data and partition into bins.
  - then one can smooth by bin means, bin median, bin boundaries etc.

- Regression
  - smooth by fitting
- Clustering
  - Detect and remove outliers

Simple Discretization Methods (Binning)

- Equal-width (distance partitioning)

  * Divides the range into N intervals of equal size: uniform grid.
  * if A and B are lowest and highest values of the attribute, the width of intervals be: $W = (B-A)/N$
  * The most straightforward, but outliers may dominate presentation.
  * Skewed data is not handled well.

- Equal-Depth (frequency) partitioning:
  * Divides the range into N intervals, each containing approximately same number of samples.
  * Good data scaling
  * Managing categorical attributes can be tricky.

Example

D = {4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34}

- partition into equal frequency (depth) bins
  - Bin1: 4, 8, 9, 15
  - Bin2: 21, 21, 24, 25
  - Bin3: 26, 28, 29, 34

- Smoothing by bin means:
  - Bin 1: 9, 9, 9, 9 mean (4, 8, 9, 15) ↦ 9
  - Bin 2: 23, 23, 23, 23 mean (21, 25, 21, 24) ↦ 23 problem: variance in bins
  - Bin 3: 29, 29, 29, 29 mean (26, 28, 29, 34) ↦ 29

- Smoothing by bin boundaries
  - Bin1: 4, 4, 4, 15
  - Bin2: 21, 21, 25, 25
  - Bin3: 26, 26, 26, 34

3-Differences and Ratios

- Differences and ratio transformations are not only useful for output features but also for inputs.

- They can be used as changes in time for one feature as a composition of different input features. (BMI)

Missing Data

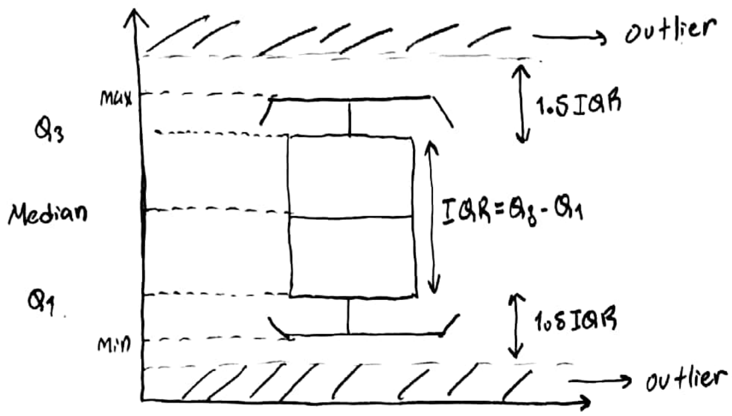
- Replace all missing values with a single global constant
- Replace a missing value with its feature mean.
- Replace a missing value with its feature mean for the given class.
- If the more than %50 of a feature is missing, then we can remove this feature.

Outlier Analysis

+ Quartiles, outliers and boxplots
  - Quartiles: Q_1 (25^{th} percentile), Q_3 (75^{th} percentile)
  - Inter-quartile range: $IQR = Q_3 - Q_1$
  - Five Number Summary: min, Q_1, M, Q_3, max $\longmapsto$ Q_1 = median (M, min), Q_3 = Median (M, max)
  - Boxplot: ends of the box are the quartiles, median is marked, whiskers, and plot outlier individually.

  - Outlier: usually, a value higher/lower than $1.5 \times IQR$

  - A common rule for identifying suspected outliers is to single out values falling at least $1.5 \times IQR$ above the third quartile or below the first quartile.



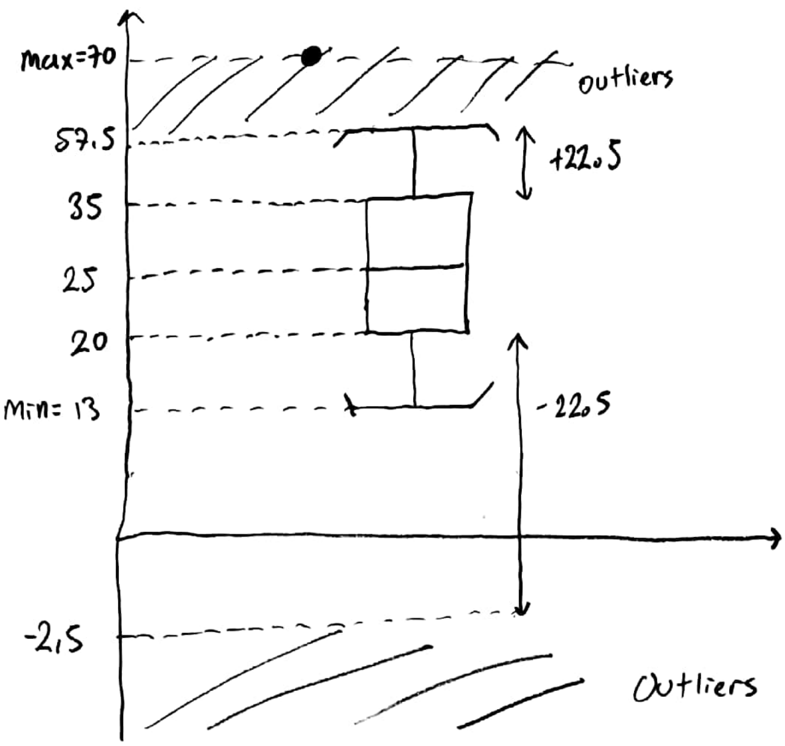
Example

Values = {13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70}

Min	Q_1	M	Q_3	max
13	20	25	35	70

$IQR = Q_3 - Q_1 = 15$

$1.5\,IQR = 22.5$



Data Reduction

1- Feature Selection

Based on the knowledge of the application domain and the goals of the mining efford, the human analyst may select a subset of the features found in the initial dataset. (Finding a subset of features)

- Comparison of means and variances.
  - The weakness of this approach is that the distribution of feature is unknown. If it is assumed to be Gaussian, it works well but it is a poor assumption.
  - Next equations formalize the test, where A and B are sets of feature values measured for two different classes, and n_1 and n_2 are the corresponding number of samples

$$SE(A-B) = \sqrt{var(A)/n_1 + var(B)/n_2}$$

$$TEST: |mean(A) - mean(B)| / SE(A-B) > \text{threshold value}$$

We want test to be large enough.
If it is, then this feature is distinguishing between A and B classes.

X	y	C
0.2	0.9	B
0.3	0.7	A
0.6	0.6	A
0.5	0.5	A
0.7	0.7	B
0.4	0.9	B

→

X	y	Class
0.3	0.7	A
0.6	0.6	A
0.5	0.5	A
0.2	0.9	B
0.7	0.7	B
0.4	0.9	B

$mean(X_A) = 0.4667$ | $Var(X_A) = 0.023$

$mean(X_b) = 0.4333$ | $Var(X_B) = 0.633$

$$TEST(X) = \frac{|0.4667 - 0.4333|}{\sqrt{\frac{0.023}{3} - \frac{0.633}{3}}} = 0.0735$$

$Mean(y_A) = 0.6$ | $var(y_A) = 0.01$

$mean(y_B) = 0.833$ | $var(y_B) = 0.013$

$$TEST(y) = \frac{|0.6 - 0.833|}{\sqrt{\frac{0.01}{3} + \frac{0.013}{3}}} = 2.66$$

feature reduction

- feature selection
  - supervised
    - Selection with class mean and variance
  - unsupervised
    - entropy measure for ranking features
- feature composition (PCA)

Entropy Measure For Ranking Features

The algorithm is based on a similarity measure S that is in inverse proportion to the distance D between two n-dimensional samples.

- The distance measure is small for close samples.

$$S_{ij} = \exp(-\alpha D_{ij})$$

where D_{ij} is the distance between samples x_i and x_j and α is

$$\alpha = -(\ln 0.5)/D$$

Where D is the average distance among samples in dataset.

- Normalized euclidean distance measure is used to calculate D_{ij}:

$$D_{ij} = \left[\sum_{k=1}^{n} ((x_{ik} - x_{jk})/(\max_k - \min_k))^2 \right]^{1/2}$$

Where n is number of dimensions and $\max_k$ and $\min_k$ are maximum and minimum values used for normalization of the k-th dimension.

- Hamming distance for nominal variables:

$$S_{ij} = \left(\sum_{k=1}^{n} |x_{ik} = x_{jk}| \right) / n$$

- For a dataset of N samples, the entropy measure is

$$E = - \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left(S_{ij} \times \log S_{ij} + (1 - S_{ij}) \times \log (1 - S_{ij}) \right)$$

Algorithm: Entropy Measure For Ranking Features:

1- Start with the initial full set of features F.

2- For each feature $f \in F$, remove one feature f from F and obtain a subset F_f. Find the difference between entropy for F and entropy for all F_f. Let f_k be a feature such that the difference between entropy for F and entropy for F_{f_k} is minimum.

3- Update the set of features $F = F - \{F_k\}$, where $-$ is a difference operation on sets.

4- Repeat steps 2-3 until there is only one feature in F.

END

Data Reduction

- Feature Reduction
- Sample Reduction
- Value Reduction

Example

Distance based outliers $(d=3,\ p=4)$

X	Y
2	4
3	2
1	1
4	3
1	6
5	3
4	2

$$D = \begin{bmatrix} 0 & \sqrt{5} & \sqrt{10}\checkmark & \sqrt{5} & \sqrt{5} & \sqrt{10}\checkmark & \sqrt{8} \\ & 0 & \sqrt{5} & \sqrt{2} & \sqrt{20}\checkmark & \sqrt{5} & 1 \\ & & 0 & \sqrt{13}\checkmark & 5\checkmark & \sqrt{20}\checkmark & \checkmark\sqrt{10} \\ & & & 0 & \sqrt{18}\checkmark & 1 & 1 \\ & & & & 0 & 5\checkmark & 5\checkmark \\ & & & & & 0 & \sqrt{2} \\ & & & & & & 0 \end{bmatrix}$$

Sample 1 $\mapsto p=2$

Sample 2 $\mapsto p=1$

Sample 3 $\mapsto p=5$ — outliers

Sample 4 $\mapsto p=2$

Sample 5 $\mapsto p=5$

Sample 6 $\mapsto p=3$

Sample 7 $\mapsto p=2$

$$Var(x) = \frac{\sum_i (x-\bar{x})^2}{n-1} \qquad std(x) = \sqrt{\frac{\sum_i (x-\bar{x})^2}{n-1}}$$

$$COV_{i,j} = \frac{\sum_k (x_{k,i} - \bar{x_i})(x_{k,j} - \bar{x_j})}{m-1}$$

for n feature data

$COV_{i,j} \in \mathbb{R}^{n \times n}$ and $COV_{i,j} = COV_{j,i}$

Principal Component Analysis

1- Subtract the mean (zero mean) $\mapsto (x,y) - (\bar{x},\bar{y})$

2- Calculate the covariance matrix

3- Calculate the eigen vectors and eigenvalues of covariance matrix.

4- [Data adjust] $\cdot$ $[U_1 \quad U_2]$ = [Data transform]

Example

1- $[\text{Data}]_{150\times4}$

2- $[\text{Adjust Data}]_{150\times4}$

3- $[\text{covariance matrix}]_{4\times4}$

4- $\begin{matrix} \lambda_1 & \lambda_2 & \lambda_3 & \lambda_4 \\ U_1 & U_2 & U_3 & U_4 \end{matrix}$ reorder $\lambda_1 > \lambda_2 > \lambda_3 > \lambda_4$

5- $\frac{\lambda_1}{\sum_i \lambda_i} \geq 0.98?$, $\frac{\lambda_1 + \lambda_2}{\sum_i \lambda_i} \geq 0.98?$

6- $[\text{Adjust Data}]_{150\times4} \cdot [U_1 \quad U_2]_{4\times2} = [\text{Data transform}]_{150\times2}$

Metrics

		Predicted Class	
		1	0
Actual Class	1	TP	FN
	0	FP	TN

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP}$$

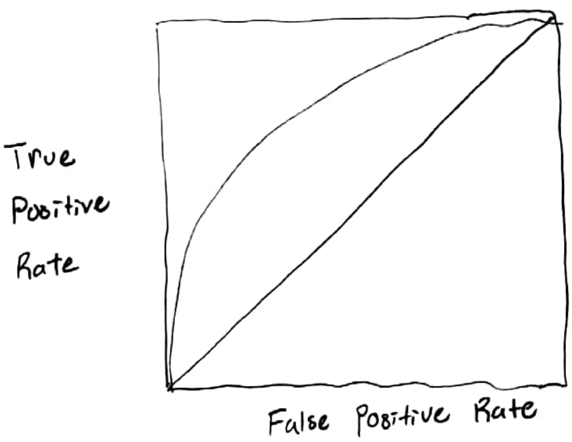
$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN} = \text{sensitivity}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

ROC Curve

Need continuous predictions!



$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

Class Imbalanced Problem

- class-based ordering
- cost-sensitive classification
- sampling-based approaches

$C(i, j)$: cost of misclassifying class i example as class j

$$\text{cost} = \sum C(i,j) \times f(i,j)$$

we want cost low.

Data Splitting

1- Resubstitution Method
2- Holdout Method
3- Leave-one-out Method
4- Rotation Method (K-fold cross validation)
5- Bootstrap Method

Binary Similarities

		j sample	
		0	1
i sample	0	a	b
	1	c	d

Simple Matching Coeff (SMA) = $\frac{a+d}{a+b+c+d}$

Jaccard = $\frac{d}{b+c+d}$

K-Means Algorithm

N datapoints into K disjoint subsets S_J

Minimize:

$$J = \sum_{J=1}^{K} \sum_{n \in S_J} |X_n - \mu_J|^2$$

Where X_n represents nth data point and μ_J is geometric centroid of S_J

1- Randomly initialize K centroids $\mu_1, \ldots, \mu_K$

2- for all $X_J \in D$:

  for all $i \in [1, K]$:

    compute $\delta(X_J, \mu_i)$

    assign X_J to closest centroid, $C_i^* = \arg\min \delta(X_J, \mu_i)$

  end for

end for

3- for all $i \in [1, K]$:

$$\mu_i' = \frac{1}{|C_i|} \sum_{J=1}^{|C_i|} X_J$$

4- if $\sum_i |\mu_i' - \mu_i| < \epsilon$ then:

    Stop

  else:

    go to 2

end if

Hiecarchical Clustering

Agglomerative or divisive

- Agglomerative → bottom up (merging)
- divisive → top down (splitting)

Agglumerative

- Linkage Methods
  - Single Linkage
  - Complete Linkage
  - Average Linkage
- Variance Methods
  - Ward's Method
- Centroid Methods

Dendogram : Hierarchical Clustering

Distance Between Clusters

Single linkage

Max distance

↳ complete linkage algorithm

Min distance ⟶ $d_{min}(C_i, C_j) = \min\limits_{\substack{p \in C_i \\ p' \in C_j}} |p - p'|$

Mean distance ⟶ $d_{mean}(C_i, C_j) = |m_j - m_i|$

Average distance → $d_{avg}(C_i, C_j) = \frac{1}{n_i \cdot n_j} \sum\limits_{p \in C_i} \sum\limits_{p' \in C_j} |p - p'|$